EXHIBIT 1

Class action against GitHub Copilot

Class action against GitHub Copilot

Posted Nov 10, 2022 16:26 UTC (Thu) by **bluca** (subscriber, #118303)
Parent article: Class action against GitHub Copilot

I am so glad to live in Europe, where the legislation is way ahead of the US on this matter and makes clear that such a lawsuit is absolutely bogus and nonsensical. There are explicit provisions for text and data mining for the purpose of AI being excepted from copyright laws, as it should be. In the US it has to rely on fair use - mind, this lawsuit is still bogus and nonsensical, but it means it will have to rely on the court to do the right thing, as fair use is a case-by-case affair.

The demand for money makes it even more obvious this is a malicious effort by some copyright trolls, pushing for a maximalist interpretation of the law, which would be bad news for everybody but their lawyers and wallets.

(Log in to post comments)

Class action against GitHub Copilot

Posted Nov 10, 2022 17:14 UTC (Thu) by **Rigrig** (subscriber, #105346) [Link]

Training an AI should be fine, but when it faithfully reproduces input it gets tricky:

Take the `isEven()` function from the complaint:
If anyone writes an `isEven()`, chances are it looks like a lot of other `isEven()` functions out there.
But this one is exactly the same as a textbook example, which uses *recursion for every input except `0` or `1`*. It even includes the test code from the book, including the `// -> ??` exercise comment.

Regardless if it was produced by AI or a human, that sure smells like copyright violation to me.
Which is also what a lot of people are worried about: that this will be used to blatantly violate copyrights by claiming "It was written by an AI, which means it's the product of fair use."

Reply to this comment

Class action against GitHub Copilot

Posted Nov 10, 2022 19:11 UTC (Thu) by **bluca** (subscriber, #118303) [Link]

Nah, that is called concern trolling or sealioning. Those are not real world use cases, they are fabricated for clickbait effect. Absolutely nobody who's working on something real goes around trying to recreate the inverse square root algorithm or things like that that have been doing the rounds.

What this is really used for in the real world is to take care of boilerplate and such.

Reply to this comment

Class action against GitHub Copilot

Posted Nov 10, 2022 21:05 UTC (Thu) by **ballombe** (subscriber, #9523) [Link]

How do you know that ?

Reply to this comment

Class action against GitHub Copilot

Posted Nov 10, 2022 21:17 UTC (Thu) by **bluca** (subscriber, #118303) [Link]

I use it every day, and talk to other developers who use it every day. Do you?

Reply to this comment

Class action against GitHub Copilot

Posted Nov 11, 2022 9:51 UTC (Fri) by **gspr** (subscriber, #91542) [Link]

Someone has made a machine that outputs *potentially* copyright-infringing code. Whether or not it really is, is the core of the debate, and not something that's easy to answer. What I fail to understand is how "I and all my peers just use that machine to output non-copyrightable boilerplate" is any sort of excuse.

Reply to this comment

Class action against GitHub Copilot

Posted Nov 11, 2022 10:24 UTC (Fri) by **bluca** (subscriber, #118303) [Link]

Because it's not an excuse, it's explaining how this works in the real world, outside of clickbaity articles and copyright troll lawsuit fishing for money. Because this matters, a lot.
My smartphone camera can also *potentially" output copyright-infringing pictures. My mp3 player *potentially* plays copyrighted songs. And so on - these are tools, and their main intended and common use matters, and for alleged open source supporters to side with copyright maximalists and trolls looking for a quick payday is missing the point so much that it's not even fun anymore.

Copyright maximalism is bad for us. The only reason this gains any traction is because it's done by Microsoft, if Copilot had been built by Mozilla reactions would be quite different, and that's just sad and short-sighted.

Reply to this comment

Class action against GitHub Copilot

Posted Nov 11, 2022 11:11 UTC (Fri) by **gspr** (subscriber, #91542) [Link]

I think these analogies are terrible. Your camera and your music players are indeed tools that you can use to infringe on copyright. They didn't come to you in a state infringing on copyright. The claim about Copilot, however, is that it "contains" (for some value of contains) and produces copyrighted material.

Reply to this comment

Class action against GitHub Copilot

Posted Nov 11, 2022 11:47 UTC (Fri) by **bluca** (subscriber, #118303) [Link]

No, those analogies are perfectly adequate. Copilot does not contain any infringing material, it's not a repository. And you have to intentionally make it produce exact copies, with carefully selected inputs that you need to have pre-existing knowledge of, it does not happen randomly, it takes a lot of effort. So it's exactly like a camera or a music player, in that regard. It's the user intent that causes that behaviour, not the tool itself.

Reply to this comment

Class action against GitHub Copilot

Posted Nov 11, 2022 12:14 UTC (Fri) by **gspr** (subscriber, #91542) [Link]

> Copilot does not contain any infringing material

That's the question at the heart of the conundrum, and it's a very complex one. Sure, we can probably agree that Copilot does not contain bit-for-bit copies of copyrighted material. But that's not the bar. Distributing a lossily compressed copy of a copyrighted image without permission can still be infringement. On the other hand, distributing the average value of all the pixels in said image certainly is not. The spectrum in-between is where it gets hard, and this is (in my opinion) probably where Copilot and similar fall.

Reply to this comment

Class action against GitHub Copilot

Posted Nov 11, 2022 15:32 UTC (Fri) by **Wol** (subscriber, #4433) [Link]

As a completely different example, copying material as part of a lawsuit does not infringe copyright. I can copy a work, present it to court, and copyright cannot touch me.

But if somebody else then takes my work and publishes it in a newspaper, that's not a legal document. Me putting it in a legal document did not strip copyright, it just gave ME immunity. The publisher can still get done for it.

Cheers,
Wol

Reply to this comment

Class action against GitHub Copilot

Posted Nov 11, 2022 19:55 UTC (Fri) by **bluca** (subscriber, #118303) [Link]

Of course it is complicated, but the crux is that I don't see how this tool can be defined as something that distributed copies of anything. It is very obviously not built to pick existing snippets and shove them out of the door 1:1 to users. It's not how it works in the vast majority of cases, where it builds something adapted to the surrounding environment - which is what makes it so darn beautiful and useful to use. Then there are users who go out of their way to carefully construct a surrounding environment (=> input query) using their pre-existing knowledge to intentionally cause it to spit out pre-existing snippets in order to write clickbaity articles or, in this case, look for a quick payday.

Reply to this comment

Class action against GitHub Copilot

Posted Nov 12, 2022 15:24 UTC (Sat) by **farnz** (subscriber, #17727) [Link]

It can be defined as something that distributes copies of something in exactly the same way as a human engineer can be defined as someone who distributes copies of something.

If I have, in my notebooks, details of how to do something in kernel-dialect C, and I read a snippet of code from those details then adapt it to the codebase I'm working on, then I've distributed a copy of the snippet in my notebook. If the snippet in my notebook is not protected by copyright, then this is not an issue; if it is, then I've potentially infringed copyright by copying out that snippet and adapting it.

The same applies to Copilot - its model takes the place of the engineer's notebooks and knowledge of what they can find in their notebooks, and its output is potentially infringing in exactly the same way as a human engineer's output is potentially infringing, complete with fun questions around "non-literal copying".

Reply to this comment

Class action against GitHub Copilot

Posted Nov 17, 2022 13:16 UTC (Thu) by **esemwy** (subscriber, #83963) [Link]

A human engineer can be inspired by someone else’s code. An AI has no such ability. Computers memorize, and have rules. The fact that the AI is really complicated doesn’t change that. It’s more as if they obfuscated the source and are trying to pass it off as their own.

Reply to this comment

Class action against GitHub Copilot

Posted Nov 17, 2022 14:08 UTC (Thu) by **farnz** (subscriber, #17727) [Link]

You're getting into the philosophy of what it means to be human, and missing my point at the same time.

My point is simply that if a human can infringe while doing the same thing that Copilot does, then it's absurd to say that Copilot cannot infringe because it's an AI - rather, it's reasonable to say that Copilot's ability to infringe copyright is bounded on the lower end by the degree to which a human doing the same thing can infringe copyright.

I've also provided a sketch of how a human can infringe copyright, which I can expand upon if it's not clear; unless you can demonstrate that Copilot is incapable of doing what the human does to infringe, however, you can't then claim that Copilot can't infringe where a human can.

Reply to this comment

Class action against GitHub Copilot

Posted Nov 17, 2022 15:57 UTC (Thu) by **esemwy** (subscriber, #83963) [Link]

The distinction does make a difference in this case. If a human takes a snippet of code and copies it, only changing the identifiers, my understanding is that is still a violation of copyright, and a company can't take code, compile it, and claim it's not an infringement. Copilot seems more like the latter, where the former example would be the human analog.

Reply to this comment

Class action against GitHub Copilot

Posted Nov 17, 2022 17:10 UTC (Thu) by **farnz** (subscriber, #17727) [Link]

I'm not following your reasoning here - in what way should Copilot be permitted to do something that would make Microsoft liable for the infringement if an employee did it?

Remember that I'm setting a lower bound - "if Copilot was just a communication interface to a human being at Microsoft who looked at the context sent to and then responded with a code snippet, would Microsoft be liable?". My claim is that if Microsoft would be liable in this variant on Copilot, then Microsoft are also liable if Copilot is, in fact, an "AI" based around machine learning, but that this is a one-way inference - if Microsoft would not be liable if Copilot was a comms channel, this doesn't tell you anything about whether Microsoft are liable if Copilot is actually an AI.

To summarize: my reasoning is that "an AI did it, not a human" should never be a get-out clause - it can increase your liability beyond that you'd face if a human did it, but it can never decrease it.

Reply to this comment

Class action against GitHub Copilot

Posted Nov 11, 2022 11:06 UTC (Fri) by **paulj** (subscriber, #341) [Link]

For transparency, can you state if you have any relationships with either Microsoft or GitHub (I vaguely recall you have with MS, but ICBW)?

Reply to this comment

Class action against GitHub Copilot

Posted Nov 11, 2022 13:11 UTC (Fri) by **rahulsundaram** (subscriber, #21946) [Link]

> For transparency, can you state if you have any relationships with either Microsoft or GitHub (I vaguely recall you have with MS, but ICBW)?

Bluca works for MS but is not involved with GitHub directly as I understand it. He has said before he doesn't think it is important to add such notes but I think the repeated level of participation in topics like these warrants one. It is a complex issue and it was clear from the beginning this is all going to end up in court(s). It may very well end with rulings that affect the future of such tools and even copyright in general. If you are going to come in strongly on one side (even if it happens to be coincidentally favorable to your employer which I can completely accept it is), other folks might want to take that into consideration when evaluating your opinion.

Reply to this comment

Class action against GitHub Copilot

Posted Nov 11, 2022 14:30 UTC (Fri) by **bluca** (subscriber, #118303) [Link]

Absolute drivel. This is the comments section of a news article, not a court room. Are you going to post your tax returns to show you haven't invested in any of the companies involved? No? Thought so. Wind yer heid in.

Reply to this comment

Class action against GitHub Copilot

Posted Nov 11, 2022 14:50 UTC (Fri) by **paulj** (subscriber, #341) [Link]

Aye right pal.

Reply to this comment

Class action against GitHub Copilot

Posted Nov 12, 2022 20:16 UTC (Sat) by **k8to** (subscriber, #15413) [Link]

Others seem to think it matters, myself included.

Reply to this comment

Class action against GitHub Copilot

Posted Nov 18, 2022 1:20 UTC (Fri) by **jschrod** (subscriber, #1646) [Link]

This is not drivel.

You are currently posting 20% of the comments on this article, without disclosing your affiliation.

I.e., you are a MS shill.

plonk

PS: I know a lot of folks who work at MS Research, and I'm grateful for the great work they are doing there. It is also obvious that there are some folks at MS who are doing very good work at the Linux kernel. But they are always open about their affiliation, even when commenting articles. And lwn.net is not some obscure Web site with an obscure community -- we are here at the heart of the Linux community that takes such issues seriously and discusses is with open visor.

PPS: For the record: I'm an owner of a company that is a MS partner, but my company has nothing to do with the Linux side of MS's business.

Reply to this comment

Class action against GitHub Copilot

Posted Dec 7, 2022 11:06 UTC (Wed) by **nye** (guest, #51576) [Link]

> I.e., you are a MS shill.
> *plonk*

This kind of ad hominem trolling has no place in LWN. Corbet, please for the love of god can we start seeing some temporary bans for repeat trolls like this?

Reply to this comment

Class action against GitHub Copilot

Posted Nov 11, 2022 17:52 UTC (Fri) by **paulj** (subscriber, #341) [Link]

Yes, bluca is very vocal on topics related to his (IIRC) employer MS. Hence why the association stuck in my mind.

It's almost impossible for such associations not to colour one's thinking at least a little. The enthusiasm shown for the debate by bluca is clear anyway.

Reply to this comment

Class action against GitHub Copilot

Posted Nov 11, 2022 19:47 UTC (Fri) by **bluca** (subscriber, #118303) [Link]

Except of course you say that as if you knew if for a fact, when it is very obviously factually incorrect, and one just has to go and see what my comments were for example on the whole secure-core-pc debacle to realise how nonsensical your proposition is. So you are misrepresenting reality either out of ignorance or malice. Either way, how about you stick to the facts of the matter and leave out the ad-hominems and doxxing?

Reply to this comment

Class action against GitHub Copilot

Posted Nov 14, 2022 10:29 UTC (Mon) by **paulj** (subscriber, #341) [Link]

Nothing in my comment was ad-hominem. It is widely recognised that people's associations - particularly any with tangible benefits - may at times colour their opinions of their associates, and even indirectly, friendly associates of their associates. It is a general human thing - not specific to you. It may or may not apply to you, but I - and others - would prefer to be aware of the association.

Nor did I "doxx" you. You acknowledged your employement with MS before here on LWN.

Reply to this comment

Class action against GitHub Copilot

Posted Nov 14, 2022 10:43 UTC (Mon) by **LtWorf** (guest, #124958) [Link]

> Either way, how about you stick to the facts of the matter and leave out the ad-hominems and doxxing?

Ok, here's some facts: at the moment of writing

your comments amount to the 22% of the comments

there are 40 usernames that commented, so you are clearly over represented

Reply to this comment

Class action against GitHub Copilot

Posted Nov 11, 2022 19:03 UTC (Fri) by **ballombe** (subscriber, #9523) [Link]

> I use it every day, and talk to other developers who use it every day.

Nice to know.

> Do you?
No, that is what I ask.

My issue is that there is no verifiable claim about the size of the AI model.
For all we know it can be in the petabyte size. The model could just return table of indices to a gigantic array of strings.

Github made everyone nervous by changing the TOS. They pay the price now.

Reply to this comment

Class action against GitHub Copilot

Posted Nov 11, 2022 19:34 UTC (Fri) by **bluca** (subscriber, #118303) [Link]

If it was, it wouldn't work how it does. It very clearly learns and adapts to the surrounding context - which is what makes it truly amazing to use to deal with boilerplate and repeated local patterns. I am a very lazy person, and it saves me some keystrokes when for example adding function-level unit tests - it's an automated copy-paste-search-replace that adapts to the current content. You couldn't do that with an index that returns pre-existing snippets.

Besides, some folks are reimplementing their own server + model, using the same client interface, and yes it's an actual AI model, not an index: https://github.com/moyix/fauxpilot

Reply to this comment

Class action against GitHub Copilot

Posted Nov 13, 2022 11:49 UTC (Sun) by **ballombe** (subscriber, #9523) [Link]

> If it was, it wouldn't work how it does. It very clearly learns and adapts to the surrounding context - which is what makes it truly amazing to use to deal with boilerplate and repeated local patterns. I am a very lazy person, and it saves me some keystrokes when for example adding function-level unit tests - it's an automated copy-paste-search-replace that adapts to the current content. You couldn't do that with an index that returns pre-existing snippets.

There would be two stages: first locate the relevant code snippet in the database, and then the AI would post-process the snippet to adapt it to the context.

The second step is something that AI are well suited to do and nobody is claiming it is violating copyright, except
in so far that it is obfuscating the first stage.

The whole concept of using function name to infer their implementation requires some kind of storage, from purely information-theoretic consideration if only to conserve Kolmogorov complexity.

> Besides, some folks are reimplementing their own server + model, using the same client interface, and yes it's an actual AI model, not an index: https://github.com/moyix/fauxpilot

So even if copilot is shut down, you can go about your work by using fauxpilot ? Good!

Reply to this comment

Class action against GitHub Copilot

Posted Dec 16, 2022 7:45 UTC (Fri) by **ssmith32** (subscriber, #72404) [Link]

Using someone else's boilerplate code is still a copyright violation.

Reply to this comment

Class action against GitHub Copilot

Posted Dec 18, 2022 2:25 UTC (Sun) by **anselm** (subscriber, #2796) [Link]

> *Using someone else's boilerplate code is still a copyright violation.*

That would depend on the exact circumstances. In many jurisdictions, code must exhibit a certain minimal degree of creativity to be eligible for copyright. If the boilerplate code in question is a very obvious or indeed the only sensible way of achieving a certain result in the programming language (and just a hassle to type out), then it may not fall under copyright because it is not sufficiently creative to warrant protection.

In such cases the main advantage of GitHub Copilot is probably that it is able to regurgitate the boilerplate with adapted variable names etc. But if all you're interested in is saving yourself some typing for boilerplate code that you use often, many programming editors have their own facilities to do this in a way that is a lot simpler and safer and doesn't involve referring to a ginormous proprietary search engine with a complete disregard of the legalities and etiquette of sharing code.

Reply to this comment

Class action against GitHub Copilot

Posted Nov 10, 2022 17:15 UTC (Thu) by **MarcB** (subscriber, #101804) [Link]

> There are explicit provisions for text and data mining for the purpose of AI being excepted from copyright laws, as it should be.

Why is this "as it should be"? Obviously, an AI system must not be allowed to perform any "copyright washing". Otherwise copyleft licenses would be completely undermined and any leaked, proprietary source code could be "freed" of its license.

The existing exemptions are for the purpose of the mining itself. The final result of this is then subject to a separate check. A research paper, or statistics or some abstract summary would obviously be allowed, but in this case, the output can be the literal input (minus copyright and license information). It is absolutely not clear if this is legal under any jurisdiction.

Reply to this comment

Class action against GitHub Copilot

Posted Nov 10, 2022 19:03 UTC (Thu) by **bluca** (subscriber, #118303) [Link]

> Why is this "as it should be"? Obviously, an AI system must not be allowed to perform any "copyright washing".

Because that's drivel. It is not how this works in the real world, it's completely fabricated clickbait.

Reply to this comment

Class action against GitHub Copilot

Posted Nov 10, 2022 20:41 UTC (Thu) by **MarcB** (subscriber, #101804) [Link]

> Because that's drivel. It is not how this works in the real world, it's completely fabricated clickbait.

There are examples of to happening, so it is obviously not fabricated. It might not be an issue for the users of Copilot, because most likely the risk of developers manually copying misattributed/unattributed code from the internet is much higher, but it certainly is an issue for Microsoft.

Even if the code generated by Copilot is not a verbatim copy of the input, it is clear, that an automated transformation is not enough to free code from its original copyright. The questions then would be, how it could be shown that the AI did create the output "on its own" and who carries the burden of this proof (the plaintiff would obviously unable to do so, because they cannot access the model).

In any case, my main point was that the directives exemptions are insufficient to declare such a lawsuit nonsensical in the EU. The directive uses the following definition:
"(2) ‘text and data mining’ means any automated analytical technique aimed at analysing text and data in digital form in order to generate information which includes but is not limited to patterns, trends and correlations"

Does this cover the output of source code? Maybe, but not obviously.

Reply to this comment

Class action against GitHub Copilot

Posted Nov 10, 2022 21:02 UTC (Thu) by **Wol** (subscriber, #4433) [Link]

> Even if the code generated by Copilot is not a verbatim copy of the input, it is clear, that an automated transformation is not enough to free code from its original copyright. The questions then would be, how it could be shown that the AI did create the output "on its own" and who carries the burden of this proof (the plaintiff would obviously unable to do so, because they cannot access the model).

I think it's clear - if the plaintiff can show that the Copilot code is identical to their own, and the defendant (Copilot) had access to their code, then it's up to Copilot to prove it's not a copy.

There's also the question of "who has access to the evidence" - if you possess evidence (or should possess evidence) and fail to produce it, you cannot challenge your opponents claims over it.

So yes it is a *major* headache for Microsoft.

Oh - and as for the guy who thought "everything should be licenced GPL" - there is ABSOLUTELY NO WAY Microsoft will do that. Just ask AT&T what happened when they stuck copyright notices on Unix

...

Cheers,
Wol

Reply to this comment

Class action against GitHub Copilot

Posted Nov 10, 2022 21:16 UTC (Thu) by **bluca** (subscriber, #118303) [Link]

> There are examples of to happening, so it is obviously not fabricated.

Of course it's fabricated, complainers go out of their way to get the tool to spit out what they were looking for and then go "ah-ha!", for clickbait effect, as if it meant something. Just like using one VHS with a copied movie does not mean that the VHS company is responsible for movie piracy. Or just like if google returns a search result with a torrent link for a music track it doesn't mean google is responsible for music piracy, and so on.

> In any case, my main point was that the directives exemptions are insufficient to declare such a lawsuit nonsensical in the EU. The directive uses the following definition:
"(2) ‘text and data mining’ means any automated analytical technique aimed at analysing text and data in digital form in order to generate information which includes but is not limited to patterns, trends and correlations"
> Does this cover the output of source code? Maybe, but not obviously.

Of course it covers it, that's exactly what copilot is used for: fills in patterns (boilerplate). Have you every actually used it?

Reply to this comment

Class action against GitHub Copilot

Posted Nov 11, 2022 12:22 UTC (Fri) by **gspr** (subscriber, #91542) [Link]

> complainers go out of their way to get the tool to spit out what they were looking for and then go "ah-ha!"

does not imply

> it's fabricated

or

> for clickbait effect

> Just like using one VHS with a copied movie does not mean that the VHS company is responsible for movie piracy.

If playing back a new blank VHS tape in a particular way resulted in a blurry copy of said movie, then yeah, it perhaps it would.

> Or just like if google returns a search result with a torrent link for a music track it doesn't mean google is responsible for music piracy, and so on.

I don't see how this is even comparable.

> Of course it covers it, that's exactly what copilot is used for: fills in patterns (boilerplate). Have you every actually used it?

I'm not sure it matters what it's used for by you and your peers, if it comes with an out-of-the-box ability to also do the other things. Again: this is *not* the same as "a disk drive can be used for piracy" – the difference is that Copilot already (possibly, that's the debate) contains within it the necessary information to produce the infringing material.

Reply to this comment

Class action against GitHub Copilot

Posted Nov 11, 2022 13:57 UTC (Fri) by **farnz** (subscriber, #17727) [Link]

To choose an example at one extreme, A&M Records, Inc. v. Napster, Inc. established that while there were non-infringing uses of Napster, Napster's awareness that there were infringing uses of their technology product was enough to establish liability.

And it's worth noting in this context that Napster on its own was not infringing copyright - to infringe copyright, you needed two Napster users to actively make a decision to infringe: one to make the content available, and one to request a copy of infringing content. In other words, one user had to prompt Napster to spit out what they were looking for, and even then it wouldn't do that unless another user had unlawfully supplied that content to their local copy of Napster. In contrast, if Copilot's output infringes, it only needs the prompting user to make it infringe - which doesn't bode well for Microsoft if the court determines that Copilot's output is an infringement.

Reply to this comment

Class action against GitHub Copilot

Posted Nov 11, 2022 14:39 UTC (Fri) by **bluca** (subscriber, #118303) [Link]

Napster and its users did not have a right to ingest copyrighted materials. AI developers have a right, by law (see EU Copyright Directive), to take any source material and use it to build a model, as long as it is publicly available.

Reply to this comment

Class action against GitHub Copilot

Posted Nov 11, 2022 15:07 UTC (Fri) by **farnz** (subscriber, #17727) [Link]

That's a misrepresentation both of the Napster case (where the court deemed that the user's right to ingest copyrighted materials into the system was irrelevant), and of the EU Copyright Directive, which merely says that ingesting publicly available material into your system is not copyright infringement on its own, and that the fact of such ingestion does not make the model infringing. This does not preclude a finding of infringement by the model or its output - it simply means that to prove infringement you can't rely on the training data including your copyrighted material, but instead have to show that the output is infringing.

Reply to this comment

Class action against GitHub Copilot

Posted Nov 11, 2022 14:04 UTC (Fri) by **Wol** (subscriber, #4433) [Link]

> I'm not sure it matters what it's used for by you and your peers, if it comes with an out-of-the-box ability to also do the other things.

So you think that the sale of knives, hammers, screwdrivers etc should be banned? Because they come with an out-of-the-box ability to be used for murder. Come to that, maybe banning cars would be a very good idea, along with electricity, because they're big killers.

It's not the USE that matters. All tools have the *ability* to be mis-used, sometimes seriously. Ban cameras - they take porn pictures. But if the PRIMARY use is ABuse, that's when the law should step in. Everything else has to rely on the courts and common sense.

In the UK, carrying offensive weapons in public is illegal. Yet many of my friends - quite legally - carry very sharp knives. Because they're "tools of the trade" for chef'ing.

Cheers,
Wol

Reply to this comment

Class action against GitHub Copilot

Posted Nov 11, 2022 14:19 UTC (Fri) by **gspr** (subscriber, #91542) [Link]

Sorry, my phrasing was bad. I did not mean to refer to what actions can be taken with the thing. What I'm trying to convey is that Copilot (perhaps!) contains (some representation of) the copyrighted material, and can *therefore* be used to reproduce the material.

A pen won't reproduce a copyrighted text without a human inputting missing data, even though it of course can he used to reproduce such a text with human assistance. Copilot, on the other hand, can (maybe!)

Reply to this comment

Class action against GitHub Copilot

Posted Nov 11, 2022 14:33 UTC (Fri) by **bluca** (subscriber, #118303) [Link]

It is *allowed* to ingest copyrighted materials for the models, by law. Hence it is not subject to the original license, among other things.

Reply to this comment

Class action against GitHub Copilot

Posted Nov 11, 2022 14:38 UTC (Fri) by **farnz** (subscriber, #17727) [Link]

Your "hence" does not follow from your first statement.

The law says that the act of ingestion does not itself infringe copyright, nor does the fact of ingestion make the model infringe copyright automatically. It does not, however, says that the model is not subject to the original licence if it is found to be infringing copyright, nor does it say that the output of the model is not contributory infringement.

Reply to this comment

Class action against GitHub Copilot

Posted Nov 11, 2022 14:42 UTC (Fri) by **gspr** (subscriber, #91542) [Link]

> It is *allowed* to ingest copyrighted materials for the models, by law. Hence it is not subject to the original license, among other things.

Yeah. But it's not allowed to *reproduce* that copyrighted material in a way incompatible with the original license. On one extreme, ingesting the material to produce, say, the parity of all the bits involved, is clearly not "reproduction" - and so is OK. On the other extreme, ingesting it and storing it perfectly in internal storage and spitting it back out on demand, clearly is "reproduction" - and surely not OK.

As I see it, the whole debate is about where between those extremes Copilot falls.

I'm not claiming to have the right answer. In fact, I don't even think I have _a_ answer. But I object to your sweeping statements about this seemingly being an easy and clear case.

Reply to this comment

Class action against GitHub Copilot

Posted Nov 14, 2022 9:28 UTC (Mon) by **geert** (subscriber, #98403) [Link]

> [...] aimed at analysing text and data in digital form in order to generate information which includes but is not limited to patterns, trends and correlations"

"patterns, trends, and correlations". For code, that would be reporting e.g. that 37% of all code that needs to sort something resort to quicksort, instead of reproducing a perfect copy of the source code of your newly-developed sorting algorithm released under the GPL.

Yeah, the "is not limited to" might be considered a loophole, but I guess anything that doesn't follow the spirit would be tossed out...

Reply to this comment

Class action against GitHub Copilot

Posted Nov 11, 2022 17:42 UTC (Fri) by **mathstuf** (subscriber, #69389) [Link]

I eagerly await Microsoft's addition to Copilot's training set of the Windows and Office codebases if there's no such issue.

Reply to this comment

Class action against GitHub Copilot

Posted Nov 11, 2022 19:36 UTC (Fri) by **bluca** (subscriber, #118303) [Link]

Those are not hosted on Github (not even in the private section) but in a completely separate pre-existing git forge, so I'm afraid you'll be waiting for a long time

Reply to this comment

Class action against GitHub Copilot

Posted Nov 11, 2022 20:07 UTC (Fri) by **mathstuf** (subscriber, #69389) [Link]

So? If there's no worry about contributory infringement, why not train on it? Why limit yourselves to public code and not any code Microsoft has access to?

Reply to this comment

Class action against GitHub Copilot

Posted Nov 11, 2022 20:44 UTC (Fri) by **Cyberax** (★ **supporter** ★, #52523) [Link]

Danger of accidental leaks of secret credentials hard-coded in config files/code. Which should not happen, but often does in private code.

Please note, that it's not a question of copyright.

Reply to this comment

Class action against GitHub Copilot

Posted Nov 11, 2022 21:39 UTC (Fri) by **bluca** (subscriber, #118303) [Link]

Because it's built by Github, on Github? It's also not scraping Gitlab instances and so on. Also, believe me, just trying to get access to those instances would be such a major PITA that anybody sane would just give up, leave and go fishing. There's nothing to gain anyway, so why bother?

Reply to this comment

Class action against GitHub Copilot

Posted Nov 12, 2022 3:12 UTC (Sat) by **pabs** (subscriber, #43278) [Link]

Software Heritage have managed to ingest lots of different software sources, I'm sure Microsoft could easily manage to do the same for training GitHub Copilot, or even just get a copy from SWH.

https://www.softwareheritage.org/

Reply to this comment

Class action against GitHub Copilot

Posted Nov 12, 2022 11:29 UTC (Sat) by **bluca** (subscriber, #118303) [Link]

Why bother? Accessing gigatons of data from you own infrastructure on prem is cheap and easy. The same volume of data from third parties is going to cost an arm and a leg in bandwidth alone. Is there any evidence that spending all that money would significantly improve the quality of the models in any way?

Reply to this comment

Class action against GitHub Copilot

Posted Nov 12, 2022 15:39 UTC (Sat) by **farnz** (subscriber, #17727) [Link]

It would have been wise for Microsoft to train Copilot against their crown jewels (Office and Windows) for two reasons:

1. It makes their assertion that Copilot does not infringe anyone's copyright easier to defend if they're saying that it's safe to train it against their crown jewel codebases. The fact that MS haven't done this means that there's room to argue that they won't do it ever because of the risk of accidentally publishing parts of Windows or Office source code, and not just because of the difficulty of moving data from one business unit to another.
2. There's still a lot of people working on Windows API codebases - having Copilot trained on what are presumably the "best" codebases in the world (on average) would help those people out.

Reply to this comment

Class action against GitHub Copilot

Posted Nov 12, 2022 18:31 UTC (Sat) by **bluca** (subscriber, #118303) [Link]

1) Nah, naysayers will never, ever be happy, it would help in no way whatsoever while costing a boatload of money and effort
2) [citation needed]

Reply to this comment

Class action against GitHub Copilot

Posted Nov 12, 2022 20:33 UTC (Sat) by **mathstuf** (subscriber, #69389) [Link]

Sure, 100% satisfaction is not feasible for anything, but I think it'd make a *lot* of the skepticism subside (including mine). Why wouldn't it cost any more than Copilot already cost? Or is ingesting new code not done anymore and Copilot "frozen"? If it isn't frozen, what's the marginal cost of a few hundred million lines on top of the billions already ingested?

How the hell do you think anyone would get a citation for that? Are you saying that Microsoft doesn't have useful Win32 API usage to train on for Windows developers? Or are you saying that even Microsoft doesn't use it well enough to bother training anything on it?

Reply to this comment

Class action against GitHub Copilot

Posted Nov 12, 2022 22:00 UTC (Sat) by **Cyberax** (★ **supporter** ★, #52523) [Link]

Modern neural networks are often trained in stages, so just ingesting an additional corpus of code might indeed require retraining everything. But they'll have to do it eventually anyway.

Reply to this comment

Class action against GitHub Copilot

Posted Nov 13, 2022 16:56 UTC (Sun) by **farnz** (subscriber, #17727) [Link]

For 1, it's not about the naysayers, it's about what you can say in court to convince a judge (or jury in some US civil cases) that the naysayers are overreacting. The statement "we trained this against our crown jewels, the Windows and Office

codebases, because we are completely certain that its output cannot contain enough of our original code to infringe copyright" is a very convincing statement to a judge or jury - and even if the court finds that Copilot engages in contributory infringement of people's copyright (having seen a demo of it doing so), the court is likely to be lenient on Microsoft as a result - the fact of having trained it against their core business codebases is helpful evidence that any infringement by Copilot's output is unintentional and something Microsoft would fix, because it puts their core business at risk.

And for 2, which part do you want a citation on? That Office and Windows are a big Win32 codebase written by good developers? That people still write code for Win32? That there's boilerplate in Win32 that would be simplified with an AI assistant helping you write the code?

Reply to this comment

Class action against GitHub Copilot

Posted Nov 12, 2022 22:28 UTC (Sat) by **anselm** (subscriber, #2796) [Link]

OTOH, it could be the case that the source code for Windows and Office is so atrociously horrible that they don't want to contaminate their ML model with it -- especially if there's a chance that recognisable bits of it could leak out for everyone to see.

Reply to this comment

Class action against GitHub Copilot

Posted Nov 14, 2022 10:44 UTC (Mon) by **LtWorf** (guest, #124958) [Link]

Why didn't microsoft train copilot on windows 11 and office code?

Reply to this comment

Class action against GitHub Copilot

Posted Nov 10, 2022 20:04 UTC (Thu) by **lkundrak** (subscriber, #43452) [Link]

One thing you forgot to say: "Company that's the defendant in this case pays my bills."

You're welcome.

Reply to this comment

Class action against GitHub Copilot

Posted Nov 11, 2022 10:25 UTC (Fri) by **bluca** (subscriber, #118303) [Link]

it wasn't forgot, what value do you think that adds, precisely?

Reply to this comment

Class action against GitHub Copilot

Posted Nov 11, 2022 11:22 UTC (Fri) by **gspr** (subscriber, #91542) [Link]

Courtesy and transparency, to name two.

Reply to this comment

Class action against GitHub Copilot

Posted Nov 11, 2022 11:48 UTC (Fri) by **bluca** (subscriber, #118303) [Link]

Neither are relevant. I see that you didn't dox yourself either, so where are yours?

Reply to this comment

Class action against GitHub Copilot

Posted Nov 11, 2022 12:16 UTC (Fri) by **gspr** (subscriber, #91542) [Link]

I have no relationships with either of the parties in this case.

Reply to this comment

Class action against GitHub Copilot

Posted Nov 11, 2022 14:28 UTC (Fri) by **bluca** (subscriber, #118303) [Link]

So you say. Post your tax returns, including any buying/selling of shares, going back 10 years. You wouldn't want to appear not to be "transparent", would you now?

Reply to this comment

Class action against GitHub Copilot

Posted Nov 11, 2022 14:59 UTC (Fri) by **lkundrak** (subscriber, #43452) [Link]

I suggest you stop commenting for a bit and take some time to familiarize yourself with the levels of civility that's usual in LWN comment sections. It could help you make your point without making a complete fool of yourself. You don't seem to have noticed you're doing that, and it's just painful to watch.

Reply to this comment

Class action against GitHub Copilot

Posted Nov 11, 2022 16:30 UTC (Fri) by **bluca** (subscriber, #118303) [Link]

It takes a remarkable dose of creativity to start talking about "civility" when your only contributions so far have been ad-hominems and doxxing

Reply to this comment

Civility

Posted Nov 11, 2022 16:32 UTC (Fri) by **corbet** (editor, #1) [Link]

Speaking of civility, I think that this branch of the conversation has gone far enough - and beyond. Can we retire it here please?

Reply to this comment

Class action against GitHub Copilot

Posted Nov 12, 2022 21:52 UTC (Sat) by **vetse** (subscriber, #143022) [Link]

I don't get exactly why some folks it as some axiomatic truth that companies should be able to just vacuum up any and all public data of any kind and use it for training their ML models. I'm generally for information being freely available, but most AI-related projects I've seen (Copilot included) have left a very sour taste in my mouth that make me think the creators aren't at all considering any of the consequences of what they've made.

Reply to this comment

Class action against GitHub Copilot

Posted Nov 14, 2022 10:53 UTC (Mon) by **kleptog** (subscriber, #1183) [Link]

This gets to the heart of what the recent EU Copyright Directive is trying to achieve in this area. Large companies with lots of money are going to hoover up anything publicly available anyway. If the legal status isn't totally clear, they can just throw money at the problem. On the other hand, researchers and students were getting threatened with lawsuits when they were training models on data freely accessible on the internet. Additionally, institutions like the Internet Archive and public libraries trying to archive for the future were also being threatened.

So the likely end result was that big companies with lots of money get to make new models on lots of data, while start-ups, researchers and students who are working on the next generation of technologies in this area are stymied by possible lawsuits. This was deemed undesirable.

The chosen solution is to allow model training on any publicly available data for research and training purposes. And organisations that publish online can opt-out (in a machine readable fashion) from being used in machine learning. It doesn't say anything about the copyright status of the output of the models.

Of course, it's only a directive, so you're relying on the member states to properly implement this. But it's better than it was.

Ref: https://eur-lex.europa.eu/eli/dir/2019/790/oj

Reply to this comment

Class action against GitHub Copilot

Posted Nov 18, 2022 15:44 UTC (Fri) by **nim-nim** (subscriber, #34454) [Link]

Also, the EU does not follow common law. If the agreed upon consensus proves unworkable (for example, if some foreign mega-corporation used its quasi-monopoly in code hosting, to appropriate other people's work using a model it was the only one in position to create) the law can always be changed.

It is *sooo* refreshing to live in a legal system where past mistakes are not set in stone.

Reply to this comment

Class action against GitHub Copilot

Posted Nov 18, 2022 19:50 UTC (Fri) by **Wol** (subscriber, #4433) [Link]

Well, Parliament can always overturn Common Law.

Common Law is - certainly in its origin - just people asking judges to settle disputes. It just solidifies to stone as in "this is what seems right".

And then if it seems appropriate Parliament can come along, pass Statute Law, and toss the whole Common Law structure into the bin.

Although if the Judges think it unfair they can gut the Statute - it does happen ...

Cheers,
Wol

Reply to this comment

Class action against GitHub Copilot

Posted Nov 14, 2022 10:23 UTC (Mon) by **LtWorf** (guest, #124958) [Link]

As per other occasions, it would be nice if you disclosed that you work for microsoft.

> I am so glad to live in Europe, where the legislation is way ahead of the US on this matter and makes clear that such a lawsuit is absolutely bogus and nonsensical.

I disagree.

The thing that was ruled ok was not for generating, and certainly not for verbatim copy paste. This is different and requires a separate ruling.

> The demand for money makes it even more obvious this is a malicious effort

There has to be a demand for money, or they would be saying there was no harm done and microsoft should continue to do whatever it's doing. But the fact that there is a great number of people complaining about this online tells us that they are feeling wronged… and a court might decide just how wronged they all are.

Reply to this comment